A FORMALIZATION OF FLOATING POINT NUMERIC BASE CONVERSION

David W. Matula

Washington University

Prepared for:

Advanced Research Project Agency
Public Health Services

March 1970

Unclassified

Security Classification

AD 771746

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY *(Corporate author)*

Computer Systems Laboratory
Washington University,
St. Louis, Missouri

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3 REPORT TITLE

A Formalization of Floating Point Numeric Base Conversion

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Interim

5 AUTHOR(S) *(First name, middle initial, last name)*

David W. Matula

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
March, 1970	24	8

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
(1) DOD(ARPA) Contract SD-302 (2) NIH(DRFR) Grant No. 00396 b. PROJECT NO. (1) ARPA Project Code No. 5880 c. Order No. 655 d.	Technical Report No. 17
	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10 DISTRIBUTION STATEMENT

Distribution of this document is unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	ARPA – Information Processing Techniques, Washington, D.C. NIH., Div. of Research

13 ABSTRACT

The process of converting arbitrary real numbers into a floating point format is formalized as a mapping of the reals into a specified subset of real numbers. The structure of this subset, the set of n significant digit base β floating point numbers, is analyzed and properties of conversion mappings are determined. For a restricted conversion mapping of the n significant digit base β numbers to the m significant digit base δ numbers the one-to-one, onto, and order preserving properties of the mapping are summarized. Multiple conversions consisting of a composition of individual conversion mappings are investigated and some results on the invariant points of such compound conversions are presented. The hardware and software implications of these results with regards to establishing goals and standards for floating point formats and conversion procedures are considered.

DD FORM 1 NOV 68 1473 REPLACES DD FORM 1473, 1 JAN 64, WHICH IS OBSOLETE FOR ARMY USE.

Unclassified

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
floating point numbers						
base conversion						
significant digits						
"equivalent digits" formula						
round off error						
mixed base computational environments						
accumulated error of successive conversions						

A FORMALIZATION OF FLOATING POINT NUMERIC BASE CONVERSION

David W. Matula

TECHNICAL REPORT NO. 17

March, 1970

Computer Systems Laboratory
Washington University
St. Louis, Missouri

The process of converting arbitrary real numbers into a floating point format is formalized as a mapping of the reals into a specified subset of real numbers. The structure of this subset, the set of n significant digit base β floating point numbers, is analyzed and properties of conversion mappings are determined. For a restricted conversion mapping of the n significant digit base β numbers to the m significant digit base δ numbers the one-to-one, onto, and order preserving properties of the mapping are summarized. Multiple conversions consisting of a composition of individual conversion mappings are investigated and some results on the invariant points of such compound conversions are presented. The hardware and software implications of these results with regards to establishing goals and standards for floating point formats and conversion procedures are considered.

ABSTRACT

The process of converting arbitrary real numbers into a floating point format is formalized as a mapping of the reals into a specified subset of real numbers. The structure of this subset, the set of n significant digit base β floating point numbers, is analyzed and properties of conversion mappings are determined. For a restricted conversion mapping of the n significant digit base β numbers to the m significant digit base δ numbers the one-to-one, onto, and order preserving properties of the mapping are summarized. Multiple conversions consisting of a composition of individual conversion mappings are investigated and some results on the invariant points of such compound conversions are presented. The hardware and software implications of these results with regards to establishing goals and standards for floating point formats and conversion procedures are considered.

TABLE OF CONTENTS

LIST OF FIGURES

A FORMALIZATION OF FLOATING POINT NUMERIC BASE CONVERSION

I. INTRODUCTION AND SUMMARY

The necessity of base conversion of numeric data during some stages of computation on a digital computer is a de facto component of practical numeric computation that must be recognized in any complete analysis of digital computation. On the hardware level the trend towards establishment of computer networks with possibly differently based machines and, on the software level, the mixed-base flexibility inherent in the PL/1 language specifications, both suggest that internal data of certain jobs may be necessarily subjected to multiple conversions before job termination. Thus, references to a purportedly constant floating point datum occurring at different points during program execution might encounter altered datum values. Both hardware and software designers must recognize this problem, and each can benefit from the fundamental principles obtained by considering base conversion as a mathematical transformation. In this report we shall follow the notation of our previous work[1-4], integrating the mainstream of results from those articles with a general formal development of conversion, and providing new results particularly in the area of multiple conversions.

A fundamental analysis of base conversion is concerned first with determination of the theoretical limitations inherent in any implementation of base conversion. The actual algorithmic mechanics of any theoretically realizable conversion procedure is then a secondary (albeit nontrivial[5])problem which will not concern us in this article. The formalization we introduce provides the vehicle for studying the properties and recognizing the inherent anomalies of the conversion process, which must necessarily then guide the performance specifications for floating point representations and base conversions at both hardware and software levels.

Converting integer and fixed point data to an "equivalent" differently based number system is generally achieved by utilizing essentially $\log_\delta \beta$ times as many digits in the new base δ as were present for representing numbers in the old base β system. This simplified notion of equivalence does not extend to the conversion of floating point systems. Actually, conversion between floating point number systems introduces subtle difficulties peculiar to the structure of these systems so that no such convenient formula for equating the "numbers of significant digits" is even meaningful. Thus, our formalization of floating point base conversion is preceeded in section 2 by a careful analysis of floating point number systems.

Following our previous work[1-4], a system of floating point numbers of n significant digits to the base β is characterized as a *significance space*, S_β^n, and in theorem 1 the number of elements in S_δ^m relative to the number of elements in S_β^n within a specified interval is shown to converge to $((\delta-1)\delta^{n-1}/((\beta-1)\beta^{n-1}))\log_\delta \beta$ as the interval grows to include the whole real line. This relative density of "total membership" of two significance spaces provides only a gross comparison of the two number systems, for actually there is considerable local variation in the relative density. The *gap function*[1], $\Gamma_\beta^n(x)$, is defined as the relative difference between nearest

neighbors of S_β^n at x. A comparison of the graphs of Γ_β^n and Γ_δ^m provides more insight into the comparability of two differently based floating point systems than any simplified "equivalent digit" formula.

Having characterized floating point number systems and the gap function, the conversion of the real numbers into S_β^n both by rounding and by truncation procedures are then formalized in section 3. The order preserving properties of these conversion mappings are detailed, and the Base Conversion Theorem[2] is stated, which gives the necessary and sufficient conditions for a conversion mapping from S_β^n to S_δ^m to be (1) one-to-one and (2) onto.

The important problems associated with multiple conversions of a datum are analysed in section 4. The composition of repeated rounding and/or truncation conversions is termed a *compound conversion*, and the associated *invariant points* (i.e., the points mapped into themselves) of such a compound conversion mapping are analysed. For a compound truncation conversion it is shown that the only invariant points of the mapping are the numbers common to all of the significance spaces involved. Thus, considerable importance must be attached to the intersection of significance spaces, and in theorems 10 and 11 these intersections are shown to exhibit a crucial dependence on the commensurability of the bases. Specifically, significance spaces with commensurable bases will always jointly contain a common significance space. For example the 6-digit hexadecimal numbers and the 8-digit octal numbers both contain all 21-bit binary numbers. On the other hand the members common to significance spaces with incommensurable bases (e.g. binary and decimal) will be finite in number and, for cases of computational interest, these members will typically all fall in an interval much smaller than the interval range provided by current exponent ranges on digital computers.

As a consequence of the limited membership of points common to two incommensurable significance spaces, the multiple back-and-forth conversion of a "constant datum" between two incommensurable significance spaces by truncation conversion can accumulate error so as to invalidate even the leading digit of the value of this "constant datum". Practically, the process of updating a B.C.D. tape on a binary machine might well subject stored data which is never updated to multiple binary-decimal conversions, so B.C.D tape updating is very sensitive to such anomalies of compound conversions. Fortunately, we can show that under very general conditions iterated rounding conversion of a datum between two significance spaces quickly generates a stable pair of values each of which is a reasonable approximation of the initial datum. Under the stronger conditions given in the In-and-Out Conversion Theorem[3] rounding conversion through an intermediate significance space can be guaranteed to regenerate the initial datum of the original significance space.

In the presence of more than two incommensurable bases the possibility of cyclic conversions' accumulating error in a datum is shown to exist even under rounding conversion. Our final result resolves the problem of controlling the overall growth of accumulated conversion error within a mixed base computational environment by a process which standardizes a datum's value in each of the significance spaces involved.

II. FLOATING POINT NUMBER SYSTEMS

A formalization of floating point number systems must start with a characterization of the set of floating point numbers, preferably divorced from the cumbersome digit sequence representational notation. In previous articles[1-4] this set has been termed a significance space.

Definition: For the integers $\beta \geq 2$, called the *base*, and $n \geq 1$, called the *significance* (or *precision*), let the *significance space*, S_β^n, be the following set of real numbers:

$$S_\beta^n = \{b | b = k\beta^j \text{ for some integers } k, j \text{ where } |k| < \beta^n\}$$

For clarity we shall utilize the Greek letters α, β and δ to denote bases; the English letters a,b,c and d will denote elements of a significance space, i.e. the so called "floating point" numbers; the letters i,j,k,l,m,n,p and q will denote integers; and x,y and z will denote arbitrary real numbers.

For an element $b = k\beta^j \in S_\beta^n$ the actual floating point representation of b can be visualized as having the fixed point integer portion k represented by a sign and n or less digits to the assumed base β, with the exponent portion then represented by the integer j. The floating point representation just described will not in general be unique, with digit sequence realizations of b corresponding to both "normalized" and "unnormalized" forms possible in some cases. Considerations related to the non unique determination of k,j in $b = k\beta^j$ will be treated where necessary, however, our main concern is with membership of b in the set of real numbers S_β^n for which the form of representing b is irrelevant. Note that the significance space S_β^n differs from an actual floating point number system in that there is no bound on the exponent portion j of the members $b = k\beta^j \in S_\beta^n$. Thus S_β^n is actually an infinite set. Since we shall not concern ourselves with underflow and overflow problems, the significance space S_β^n is a perfectly acceptable model of a floating point number system for our purposes.

It is easy to visualize the change in the set S_β^n caused by varying the significance, since increasing n to n+1 maintains all members of S_β^n and adds β-1 new members uniformly spaced between every neighboring pair of members of S_β^n. The dependence of the membership of S_β^n on the base β is far more subtle. In practice it has been convenient to identify a non decimal floating point number system with the "appropriate" decimal based system, however we now show that such a purported equivalence glosses over certain inherrent anomalies between differently based floating point number systems.

A gross comparison of two differently based significance spaces can be obtained by determining the relative number of members of each space over a comparable range. For example the 3(significant) bit binary numbers between unity and one thousand are $1., 1.01_2 = 1.25, 1.10_2 = 1.5, 1.11_2 = 1.75, 10.0_2 = 2., 10.1_2 = 2.5, 11.0_2 = 3., 11.1_2 = 3.5, 100_2 = 4, 101_2 = 5, \ldots, 1100000000_2 = 768, 1110000000_2 = 896$; and the 1 (significant) digit decimal numbers over the same range are 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000. In figure 1 these members are indicated by tickmarks on the real line plotted on a logarithmic scale so that the log periodic nature of the spacing between floating point numbers is evident. The ratio of the numbers of members of these systems over this interval is 40/28=1.43. Now this ratio of membership density will vary with the choice of interval, however, a reasonable overall comparison of any two floating point systems can be calculated by determining the limit of such a ratio as the interval grows to include the whole real line.

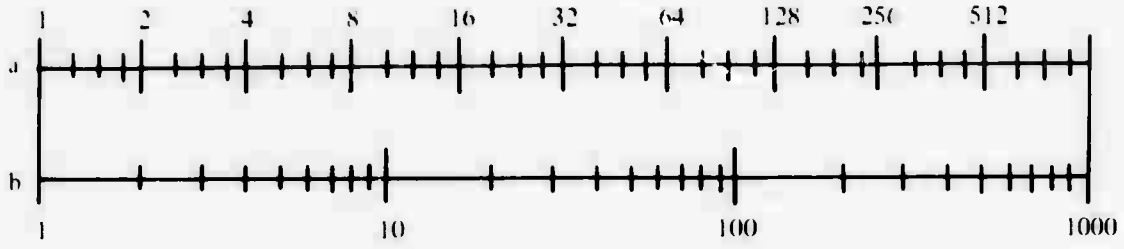

Figure 1. Tickmarks plotted on a log scale showing (a) the forty 3(significant) bit binary numbers and (b) the twenty-eight 1(significant) digit decimal numbers over the range $[1, 10^3]$.

Theorem 1: Let S_β^n and S_δ^m be any two significance spaces. Then letting $|S|$ denote the number of members of the set S,

$$\lim_{M\to\infty} \frac{\left|\left\{d\right|\ d\epsilon S_\delta^m,\ \frac{1}{M} \leq |d| \leq M\right\}\right|}{\left|\left\{b\right|\ b\epsilon S_\beta^n,\ \frac{1}{M} \leq |b| \leq M\right\}\right|} = \frac{(\delta - 1)\,\delta^{m-1}}{(\beta - 1)\,\beta^{n-1}} \log_\delta \beta \qquad (1)$$

Proof: Let $\lfloor x \rfloor$ denote the greatest integer in x. Then the closed interval $\left[\frac{1}{M}, M\right]$ may be divided into $2\left\lfloor \log_\beta M \right\rfloor$ disjoint half-open half-closed intervals of the form $\left[\beta^j, \beta^{j+1}\right)$ and two sub intervals of such intervals. Each interval $\left[\beta^j, \beta^{j+1}\right)$ contains $(\beta - 1)\beta^{n-1}$ distinct members of S_β^n. Noting that $b\epsilon S_\beta^n \Leftrightarrow b\epsilon S_{\beta}^n$ we have for $M \geq 1$,

$$\left|\left\{b\right|\ b\epsilon S_\beta^n,\ \frac{1}{M} \leq |b| \leq M\right\}\right| = 2\left|\left\{b|b\epsilon S_\beta^n,\ \frac{1}{M} \leq b \leq M\right\}\right|$$

$$= 2(2\left\lfloor \log_\beta M \right\rfloor + \epsilon)\,(\beta - 1)\beta^{n-1} \text{ where } 0 \leq \epsilon \leq 2$$

$$= 2(2\log_\beta M + \epsilon')\,(\beta - 1)\beta^{n-1} \text{ where } |\epsilon'| \leq 2$$

the latter resulting from removal of the greatest integer brackets. Finally

$$\lim_{M\to\infty} \frac{\left|\left\{d\right|\ d\epsilon S_\delta^m,\ \frac{1}{M} \leq |d| \leq M\right\}\right|}{\left|\left\{b\right|\ b\epsilon S_\beta^n,\ \frac{1}{M} \leq |b| \leq M\right\}\right|}$$

$$= \lim_{M\to\infty} \frac{2(2\log_\delta M + \epsilon_2)\,(\delta - 1)\,\delta^{m-1}}{2(2\log_\beta M + \epsilon_1)\,(\beta - 1)\,\beta^{n-1}} \qquad \text{where } |\epsilon_1|, |\epsilon_2| \leq 2$$

$$= \frac{(\delta - 1)\,\delta^{m-1}}{(\beta - 1)\,\beta^{n-1}} \log_\delta \beta$$

In the folklore on conversion there is an oft quoted notion that a decimal digit is equivalent to $\log_2 10 = 3.32\ldots$ bits. Thus a 3 digit decimal system should be only slightly less numerous than a 10 bit binary system. However, if we let $|S_\delta^m\ S_\beta^n|$ symbolically denote the limiting ratio given in equation (1), then $|S_{10}^3/S_2^{10}| = .529\ldots$, so that there are actually only 53% as many real numbers representable with 3 significant decimal digits as there are real numbers representable with 10 significant bits. Furthermore the ratio of the number of members of S_2^3 to S_{10}^1 over the range shown in figure 1 was not atypical since $|S_2^3/S_{10}^1| = 1.476\ldots$, attesting to the fact that there are about 50% *more* 3 significant bit binary numbers than 1 significant digit decimal numbers, and providing a clear contradiction to the digit = 3.32 bits rule. This anomally prevails even with more digits, since for large integral m and n chosen such that $10^m/2^n$ approaches unity,

$$\lim_{\substack{m\to\infty \\ 10^m/2^n\to 1}} |S_{10}^m / S_2^n| = \frac{9}{5}\log_{10}2 = .5418\ldots$$

The results just obtained are not inexplicable on intuitive grounds. The relation "digit = $\log_2 10$ bits" comes from an entropy argument where all different states (values) of the system (digit sequence) are distinguishable. Despite the applicability of this notion to integer and fixed point number systems, floating point systems have redundant representations for some numbers, generally resolved by normalization, so that the β^n patterns of digits associated with k for each j in $k\beta^j$ do not yield β^n new different numbers for each j. Furthermore, the collection of normalized numbers corresponding to a fixed power of the base are spread over an interval whose length depends on the base, and both of these conditions are reflected in the final form of equation (1).

Now one approach for determining a floating point "equivalent digit formula" is to equate the right hand side of (1) to unity and solve for m in terms of n, β and δ. Thus if S_δ^m is said to be *more dense* than S_β^n when $|S_\delta^m/S_\beta^n| > 1$, then from (1):

Corollary 1.1:

S_δ^m is more dense than S_β^n if and only if

$$m > n\log_\delta\beta + \log_\delta\left(\frac{\delta(\beta-1)}{\beta(\delta-1)}\right) - \log_\delta\log_\delta\beta \qquad (2)$$

Attributing meaning to a non-integral number of digits, one may propose that

$$m = n\log_\delta\beta + \log_\delta[\,\delta(\beta-1)/\beta(\delta-1)] - \log_\delta\log_\delta\beta \qquad (3)$$

is the "equivalent digit formula for floating point number systems". For binary-decimal conversion we then would have

$$\#\text{ bits} = 3.32\ldots \times (\#\text{ decimal digits}) - .884\ldots \qquad (4)$$

The variability of spacing of floating point numbers of a given S_β^n is such that the simplified formula (3) does not really provide an adequate comparison of differently based floating point systems and more attention to local magnitude dependent variability must be considered. In studying the internal structure of S_β^n note that every one of its non-zero members will have both a next largest and next smallest neighbor in S_β^n.

Definition: The *successor*, b′, of $b \epsilon S_\beta^n$, $b \neq 0$, is given by

$$b' = \min\{d | d > b, d \epsilon S_\beta^n\} \tag{5}$$

and since distinct members of S_β^n have distinct successors, b may then be referred to as the (unique) *predecessor* of b′ in S_β^n.

Now the absolute difference, b′−b, will grow with b in S_β^n, however, the relative difference is bounded.[1]

Definition: The *gap*, $\Gamma_\beta^n(x)$, in S_β^n at x is given by

$$\Gamma_\beta^n(x) = \begin{cases} \dfrac{\min\{b | b > x, b \epsilon S_\beta^n\} - \max\{b | b \leq x, b \epsilon S_\beta^n\}}{x} & \text{for } x > 0 \\ \Gamma_\beta^n(-x) & \text{for } x < 0. \end{cases} \tag{6}$$

Specifically then $\Gamma_\beta^n(a) = (a'-a)/a$ for $0 < a \epsilon S_\beta^n$. From the structure of floating point number systems it is evident that $\Gamma_\beta^n(\beta x) = \Gamma_\beta^n(x)$, so Γ_β^n will experience a log periodic behavior. For $1 \leq x < \beta$, the numerator of (6) will have the constant value β^{1-n}, so that on a log–log scale the gap function appears as a saw tooth function. In figure 2, sections of the gap functions Γ_{10}^4 and Γ_{16}^4 are illustrated. Note that the variation in the magnitude of the gap function is greater for larger bases.

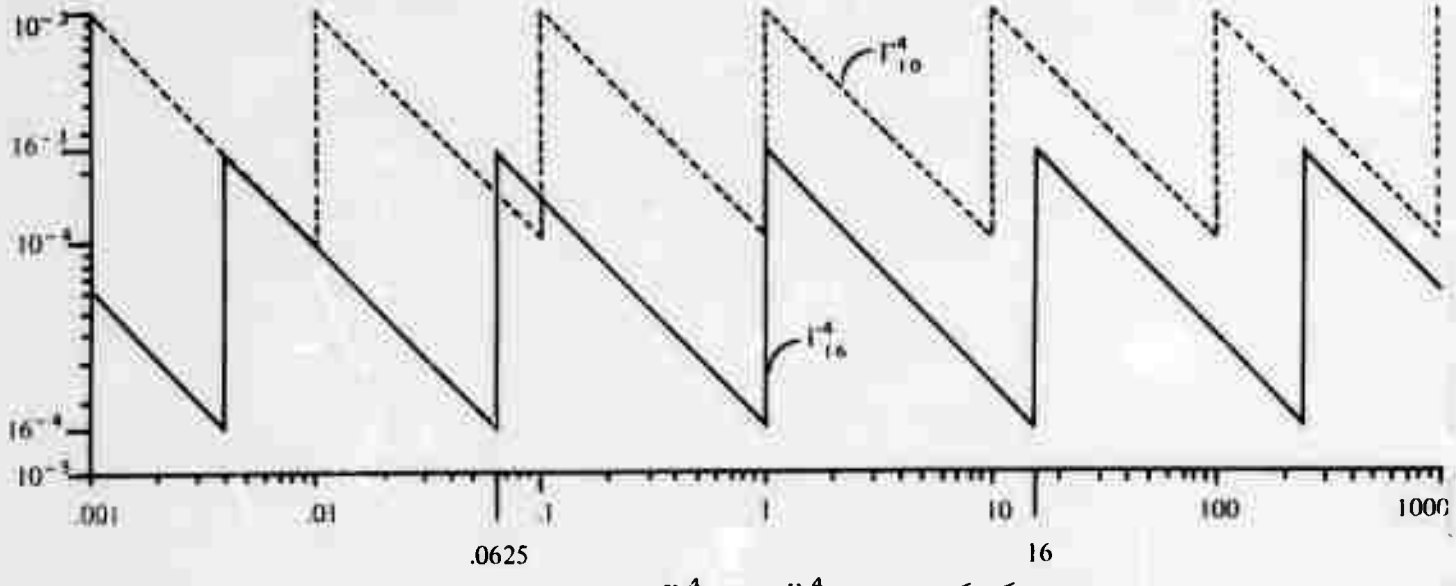

Figure 2: The gap functions $\Gamma_{10}^4(x)$ and $\Gamma_{16}^4(x)$ for $.001 \leq x \leq 1000$.

From theorem 1 we can calculate that $|S^4_{16}/S^4_{10}| = 5.66\ldots$, so that overall there are more than five times as many 4 significant digit hexadecimal numbers as 4 significant digit decimal numbers. Yet from the gap functions in figure 2 it is apparent that over the interval (.0625,.1000) there are more 4 significant digit decimal numbers that 4 significant digit hexadecimal numbers since $\Gamma^4_{10}(x) < \Gamma^4_{16}(x)$ over that interval.

From the log periodic behavior of Γ^n_β the following bounds are immediate.

Theorem 2: The function Γ^n_β attains both a minimum and a maximum value over the non-zero members of S^n_β given by

$$\begin{aligned} \min \ \left\{\Gamma^n_\beta(b) \mid b \epsilon S^n_\beta,\ b \neq 0\right\} &= 1/(\beta^n - 1) \\ \max \ \left\{\Gamma^n_\beta(b) \mid b \epsilon S^n_\beta,\ b \neq 0\right\} &= 1/\beta^{n-1} \end{aligned} \tag{7}$$

and over the non-zero reals the bounds on $\Gamma^n_\beta(x)$ are given by

$$\begin{aligned} \inf \ \left\{\Gamma^n_\beta(x) \mid x \neq 0\right\} &= 1/\beta^n \\ \max \ \left\{\Gamma^n_\beta(x) \mid x \neq 0\right\} &= 1/\beta^{n-1} \end{aligned} \tag{8}$$

Thus the gap function presents a more complete picture of the structure of a floating point number system than any "equivalent digit" notion, and simplified formulas such as (3) and (4) must be used with extreme caution in any comparison of differently based floating point number systems.

III. CONVERSION MAPPINGS

A conversion procedure determines a specific value in S_β^n for each real number x. Thus, a conversion process may be characterized as either a function or a mapping. Formalization of conversion as a mapping appears preferable since it is often useful to invert the question and refer to the set of points which map into a given element of S_β^n, and this notion is then readily available by considering the inverse mapping. Certainly any conversion mapping of the reals to S_β^n should be the identity on S_β^n and in addition it is desirable for this mapping to satisfy certain order preserving properties.

Definition: A mapping, M, of a set R′ of real numbers into the reals is

1) *Weakly order preserving (isotone[6], monotone) on R′* if $x < y \Rightarrow M(x) \leqslant M(y)$ for all $x, y \epsilon R'$

2) *Strongly order preserving on R′* if $x < y \Rightarrow M(x) < M(y)$ for all $x, y \epsilon R'$

Furthermore the mapping is said to be *weakly (strongly) order preserving* if it is weakly (strongly) order preserving on the reals.

No conversion mapping of the reals into S_β^n can be strongly order preserving, however, we can expect that a conversion process should at least be weakly order preserving in addition to being the identity on S_β^n. These latter two conditions do assure that the inverse image of any $b \epsilon S_\beta^n$ under a conversion mapping is an interval containing b. Formally we shall limit our discussion to the rounding and truncation (sometimes called chopping) conversion procedures usually encountered in computerized numeric processing.

Definition: The *truncation conversion mapping*, T_β^n, and the *rounding conversion mapping*, R_β^n, of the real numbers into S_β^n are defined for all integers $\beta \geqslant 2$, $n \geqslant 1$ as follows:

Truncation Conversion

$$T_\beta^n(x) = \begin{cases} \max \{b | b \leqslant x, b \epsilon S_\beta^n\} & \text{for } x \geqslant 0 \\ \min \{b | b \geqslant x, b \epsilon S_\beta^n\} & \text{for } x < 0 \end{cases} \tag{9}$$

Rounding Conversion

$$R_\beta^n(x) = \begin{cases} \min \left\{ b \mid \frac{b+b'}{2} > x, b \epsilon S_\beta^n \right\} & \text{for } x > 0 \\ \min \left\{ b \mid \frac{b+b'}{2} \geq x, b \epsilon S_\beta^n \right\} & \text{for } x < 0 \\ 0 & \text{for } x = 0 \end{cases} \tag{10}$$

The effects of these conversion mappings in the neighborhood of a power of the base are shown in figure 3. Note that the distinctions in the definitions of the mappings of positive and negative values are required to achieve the desired sign complementery relations:

$$T_\beta^n(-x) = -T_\beta^n(x), \; R_\beta^n(-x) = -R_\beta^n(x). \tag{11}$$

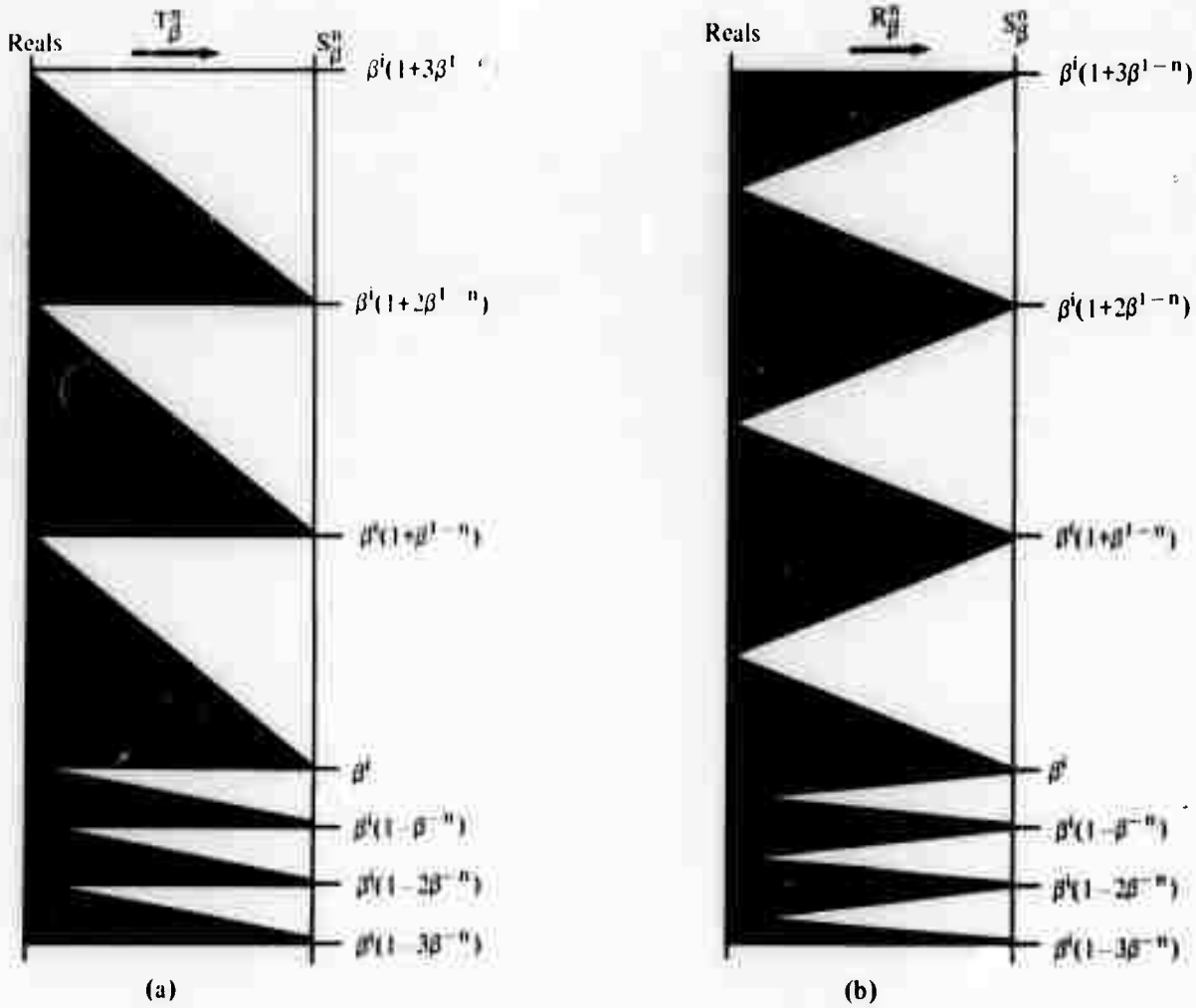

Figure 3: Conversion of the real numbers to the n(significant) digit base β numbers in the neighborhood of a power of the base by (a) truncation conversion, T_β^n, and (b) rounding conversion, R_β^n.

Note $R_\beta^n\left(\frac{b+b^-}{2}\right) = b$ for all positive $b \in S_\beta^n$, so that we have not imposed the additional symmetric rounding condition for mid-points dependent on the "parity" of b that some researchers prefer.[7] Although this refinement could be added without materially affecting our results, we prefer the definition above.

It is evident that T_β^n and R_β^n both are weakly order preserving mappings which are identities on S_β^n as desired.

In practical numeric computation we often need to convert data already expressed in floating point form to a differently based floating point form. Thus we are interested in the properties of the restricted mappings $R_\beta^n|S_\delta^m \to S_\beta^n$ and $T_\beta^n|S_\delta^m \to S_\beta^n$.

From consideration of the algorithmic mechanics of conversion there is evidently a considerable difference between binary → hexadecimal conversion and binary → decimal conversion. Since special relationships between bases do affect the properties of restricted conversion mappings there is need to characterize this important "commensurability" relationship between bases.

Definition: Let $\beta \geq 2$ be a root free integer, i.e. β has no integral i^{th} root for any i. Then the numbers $\beta, \beta^2, \beta^3, \ldots$ form a *commensurable family* of bases termed the *β-family* of bases. Two or more bases belonging to the same commensurable family of bases are *commensurable bases*. Two bases which do not belong to a common commensurable family are termed *incommensurable bases*. Furthermore, two or more significance spaces will be termed *commensurable* when their bases are commensurable.

Thus two, eight and sixteen are commensurable bases, whereas base ten is incommensurable with any member of the binary family of bases. The root free condition on β in this definition simply assures that each base is in precisely one family.

A useful equivalent characterization of commensurable and incommensurable bases avoiding explicit mention of the respective families to which the bases belong is provided in the following.

Lemma 3: The bases β and δ are commensurable if and only if $\beta^i = \delta^j$ for some non-zero integers i, j.

Corollary 3.1: The bases β and δ are commensurable if and only if $\log_\delta \beta$ is rational.

Thus, the gap functions Γ_β^n and Γ_δ^m plotted on a log-log scale will share a common period when β and δ are commensurable, as in figure 4, and otherwise will not (see figure 2).

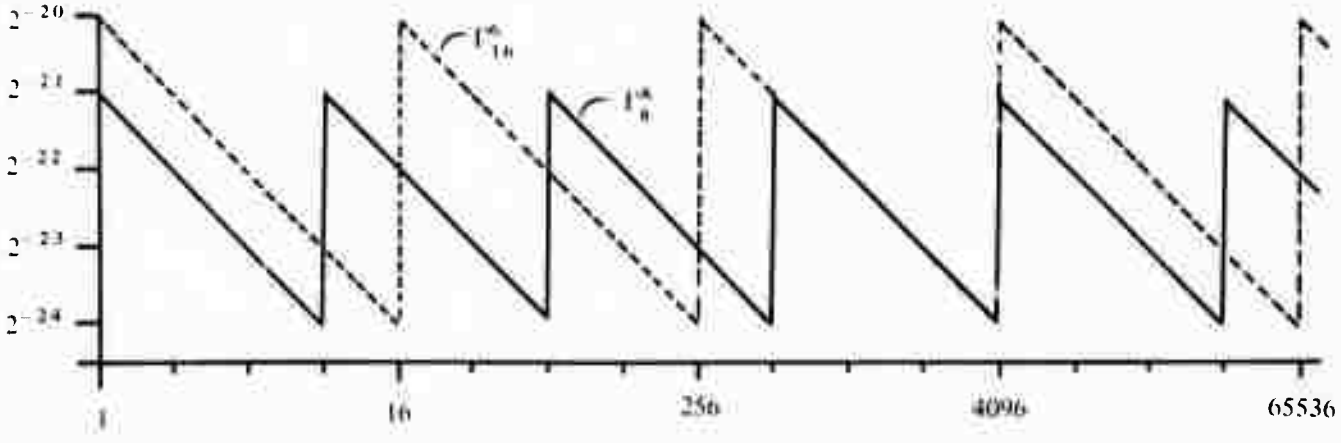

Figure 4: Sections of the gap functions Γ_{16}^6, for the 6 (significant) digit hexadecimal numbers and Γ_8^8, for the 8 (significant) digit octal numbers. Hexadecimal and octal are commensurable bases and share a common period of 2^{12} on the log scale. Note that the 6 digit hexadecimal numbers and the 8 digit octal numbers are not equivalent floating point systems.

It is readily verified that the conversion by rounding or truncation from a fixed point number system with n β ary digits to the right of the radix point to another fixed point system with m δ ary digits to the right of the radix point will be one-to-one if $m \geq n \log_\delta \beta$ and onto if $m \leq n \log_\delta \beta$. Thus a conversion between fixed point number systems must be either one-to-one or onto, however a conversion between floating point number systems need be neither one-to-one nor onto, as seen in figure 5. The necessary and sufficient conditions for rounding and truncation conversions between incommensurable significance spaces to be (1) one-to-one mappings and (2) onto mappings have been determined[2] in the Base Conversion Theorem.

Theorem 4 (Base Conversion Theorem): For incommensurable bases β and δ, the truncation (rounding) conversion mapping of S_β^n to S_δ^m, i.e., $T_\delta^m \mid S_\beta^n \rightarrow S_\delta^m \left(R_\delta^m \mid S_\beta^n \rightarrow S_\delta^m \right)$

$$\begin{aligned} &\text{i. is one-to-one if and only if } \delta^{m-1} \geq \beta^n - 1 \\ &\text{ii. is onto if and only if } \beta^{n-1} \geq \delta^m - 1 \end{aligned} \qquad (12)$$

The details of the proof of this theorem are given in [2], however, an intuitive understanding for the result can be gleamed from comparing the gap functions (see figure 2). Certainly if Γ_δ^m is uniformly less than Γ_β^n, then the mapping of S_β^n to S_δ^m should be one-to-one. Conversely, if the maximum of Γ_δ^m falls above the minimum (restricted to S_β^n) of Γ_β^n, then from a theorem of Kronecker[8] (that the integer multiples of an irrational number mod 1 are dense in the unit interval) it can be shown that for some $b \in S_\beta^n$, $\Gamma_\delta^m(b) > \Gamma_\beta^n(b)$.

In summary the essential effect of formulae (12) regarding the conversion of the initial system S_β^n to the target system S_δ^m is that to assure one-to-one conversion a digit must be sacrificed in the target system and to assure onto conversion a digit must be sacrificed in the initial system.

The conditions for one-to-one conversion guarantee another desirable property of the conversion mapping, since it is readily shown that a weakly order preserving mapping on R' is strongly order preserving on R' if and only if the mapping is one-to-one from R' to the reals.

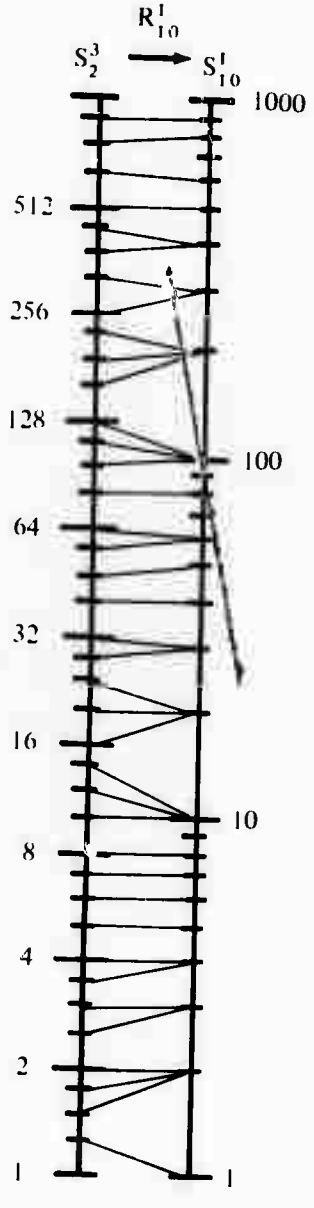

Figure 5: Conversion by rounding of the 3 (significant) bit binary numbers to the 1 (significant) digit decimal numbers over the range [1,1000], indicating that $R_{10}^1 \mid S_2^3$ is neither one-to-one nor onto.

Corollary 4.1: T_δ^m and R_δ^m are each strongly order preserving on S_β^n if and only if $\delta^{m-1} \geq \beta^n - 1$.

It is instructive to consider the implications of the Base Conversion Theorem for decimal to binary conversion. It follows that the mapping by rounding or truncation conversion of a decimal based significance space to a binary based significance space will be one-to-one (and strongly order preserving) if and only if

$$\#\text{bits} > 3.32\ldots \times (\#\text{digits}) + 1 \qquad (13)$$

and onto if and only if

$$\#\text{bits} \leq 3.32\ldots \times (\#\text{digits}) \quad 3.32\ldots \qquad (14)$$

Relating these inequalities (13, 14) to the membership density formula (1), we conclude that decimal to binary floating point conversion will be (i) one-to-one and strongly order preserving if and only if the decimal system has less than $.9\ (\log_{10}2) = .271$ times the membership of the binary system and (ii) onto if and only if the decimal system has more than $18\ (\log_{10}2) = 5.418$ times the membership of the binary system. Thus conversions between floating point systems of nearly equal density will be neither one-to-one nor onto.

The properties of decimal-binary conversion are succintly presented in figure 6. The lattice point n, m corresponds to the binary system S_2^n and the decimal system S_{10}^m. Lattice points falling to the left of the line $n = (\log_2 10)\ m+1 = 3.32m+1$ correspond to decimal to binary conversions which are one-to-one and strongly order preserving, as well as binary to decimal conversions which are onto. Lattice points falling to the right of the line $n = 3.32m\ \ 3.32$ correspond to decimal to binary conversions which are onto and to binary to decimal conversions which are one-to-one and strongly order preserving. Lattice points falling between $n = 3.32m+1$ and $n = 3.32m\ \ 3.32$ correspond to decimal-binary conversions which have none of these three properties. The equal density line $n = 3.32m\ \ .88$ separates the lattice points so that those to the left correspond to binary systems which are more dense than the decimal systems and lattice points to the right correspond to the decimal system being more dense. An increase of one unit on the n axis increases $|S_2^n/S_{10}^m|$ by a factor of 2, and a one unit increase on the m axis decreases this density ratio by a factor of 10, so $|S_2^n/S_{10}^m|$ may be easily estimated by determining the distance of the lattice point n, m from the equal density line.

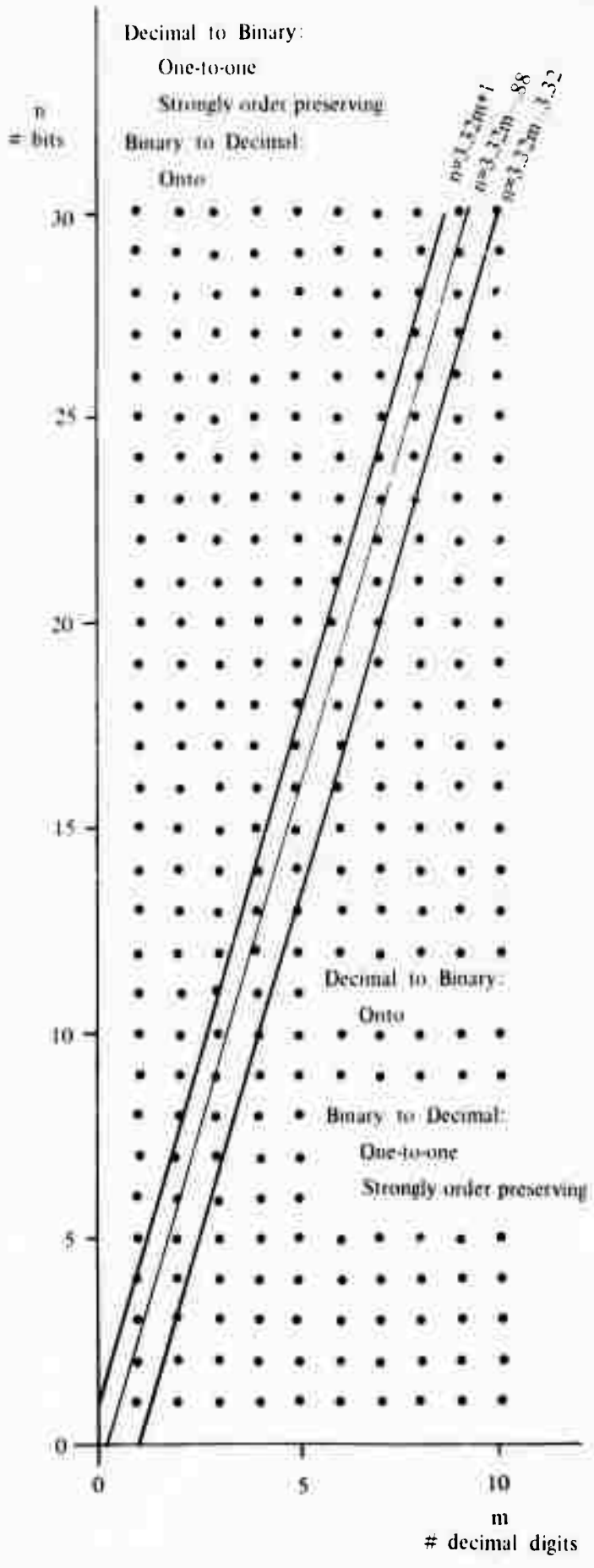

Figure 6: A summary of the properties of floating point decimal-binary base conversion.

IV. COMPOUND CONVERSIONS

The preceding section dealt only with the properties of a single conversion mapping. We have indicated previously that computing environments such as PL/I language programming, multi-computer networks and B.C.D. tape updating on a binary machine present situations where data may be subjected to multiple base conversions during overall job execution. Considerable care must be provided in such mixed base computing environments to avoid excessive accumulation of error in purportedly "constant" data. To critically analyze this problem the notion of compound conversion is formally introduced.

Definition: For all $n \geq 1$, $\beta \geq 2$,

i) T_β^n and R_β^n are *1-fold compound conversions through* S_β^n

ii) for Q a k-fold compound conversion through $S_{\beta_1}^{n_1}, S_{\beta_2}^{n_2}, \ldots, S_{\beta_k}^{n_k}$, $T_\beta^n Q$ and $R_\beta^n Q$ are *(k+1)-fold compound conversions through* $S_{\beta_1}^{n_1}, S_{\beta_2}^{n_2}, \ldots, S_{\beta_k}^{n_k}, S_\beta^n$

Furthermore Q is a *compound truncation (rounding) conversion* if all the individual conversions composing Q are truncation (rounding) conversions.

Thus $R_\beta^n R_\delta^m$ is a 2-fold compound rounding conversion through S_δ^m, S_β^n, and $R_\delta^m T_\alpha^i R_\delta^m R_\beta^n$ is a 4-fold compound conversion through S_β^n, S_δ^m, S_α^i, S_δ^m (see figure 7).

A compound conversion is a composition of mappings, and many properties of individual mappings readily carry over to compositions of such mappings. Thus the following important properties of compound conversions are immediate.

Lemma 5: If Q is a compound conversion, then $Q(-x) = -Q(x)$ for all x.

Lemma 6: Compound conversions are weakly order preserving.

An evident property of truncation conversion is that the magnitude of $T_\beta^n(x)$ is never larger than the magnitude of x. Thus truncation conversion performs a contraction of the reals towards zero (see figure 3a).

Definition: The function M:Reals→Reals is a *contraction* if M(x) has the same sign as x and $|M(x)| \leq |x|$ for all x.

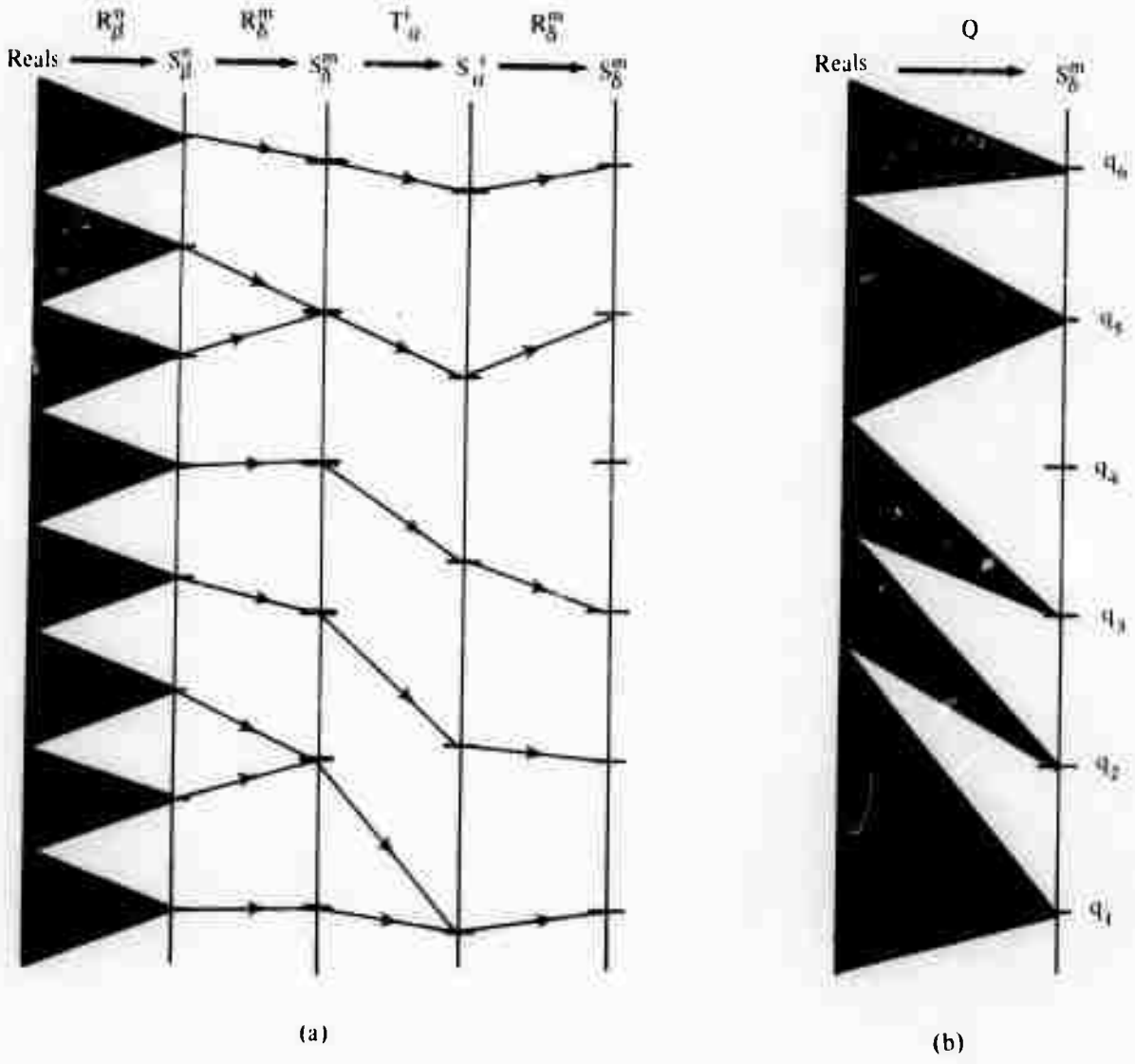

Figure 7: The composition of a compound conversion from successive conversions:
(a) the sequence of conversions R_β^n, R_δ^m, T_α^i, R_δ^m;
(b) the 4-fold compound conversion $Q = R_\delta^m T_\alpha^i R_\delta^m R_\beta^n$.

Clearly a composition of contractions is a contraction.

Lemma 7: Compound truncation conversions are contractions.

The mappings R_β^n and T_β^n are identities on their image space, S_β^n, however $T_2^4\ T_{10}^2\ (1.8) = 1.75$ and $T_2^4\ T_{10}^2\ (1.75) = 1.625$, so that members of the image space of a compound conversion Q are not necessarily carried into themselves by the mapping Q. For the compound conversion Q illustrated in figure 7 note that q_1, q_5 and q_6 are mapped into themselves by Q, whereas q_2 and q_3 are points of the image space of Q which are not mapped into themselves.

In general the points mapped into themselves by a compound conversion may be difficult to determine but they are none the less important, since the weakly order preserving property of compound conversions assures us that no point in an interval between two such invariant points of a compound conversion can be mapped outside that interval by that compound conversion.

Formally points x such that M(x) = x are generally referred to[6] as fixed points of the mapping M. However, to avoid confusion with the computational notion of "fixed point numbers", we shall refer to fixed points of mappings as invariant points.

Definition: Let M be a mapping of the reals to the reals. Then the *invariant set of* M, denoted $f(M)$, is given by

$$f(M) = \{x \mid M(x) = x\} \tag{15}$$

and for $x \epsilon f(M)$ we say that x is an *invariant point* of the mapping M.

Our first result characterizing the invariant sets of compound conversions states that any element common to all significance spaces through which a compound conversion passes is an invariant point of that compound conversion.

Lemma 8: If Q is a k-fold compound conversion through $S_{\beta_1}^{n_1}, S_{\beta_2}^{n_2}, \ldots, S_{\beta_k}^{n_k}$, then $\bigcap_{i=1}^{k} S_{\beta_i}^{n_i} \subset f(Q)$.

Proof: For any $x \in \bigcap S_{\beta_i}^{n_i}$, each of the conversions composing Q maps x into itself since $R_{\beta_i}^{n_i}$ and $T_{\beta_i}^{n_i}$ are identities on $S_{\beta_i}^{n_i}$ for i=1, . . . , k. Thus Q(x) = x, and $x \in f(Q)$.

Furthermore, it is now shown that if Q is a compound truncation conversion, then the points in the intersection of the $S_{\beta_i}^{n_i}$ are the only invariant points of the mapping.

Theorem 9: If Q is a k-fold compound truncation conversion through $S_{\beta_1}^{n_1}, S_{\beta_2}^{n_2}, \ldots, S_{\beta_k}^{n_k}$, then $f(Q) = \bigcap_{i=1}^{k} S_{\beta_i}^{n_i}$.

Proof: For $Q = T_{\beta_k}^{n_k}\ T_{\beta_{k-1}}^{n_{k-1}} \ldots T_{\beta_1}^{n_1}$, by lemma 8, $\bigcap_{i=1}^{k} S_{\beta_i}^{n_i} \subset f(Q)$. Now assume $x \notin \bigcap_{i=1}^{k} S_{\beta_i}^{n_i}$, and let us show $Q(x) \neq x$. From sign complimentarity of Q (lemma 5), $x > 0$ may be assumed. Let j be some index such that $x \notin S_{\beta_j}^{n_j}$. Compound truncation conversions are contractions and truncation conversion is weakly order preserving so that,

$$x > T_{\beta_j}^{n_j}(x) \geq T_{\beta_j}^{n_j}\ (T_{\beta_{j-1}}^{n_{j-1}} \ldots\ T_{\beta_1}^{n_1}(x)) \geq Q(x)$$

thus $x \neq Q(x)$, proving the theorem.

Lemma 8 and theorem 9 show the importance of the intersection of significance spaces with regards to determining the invariant points of compound conversion mappings. In turn the character of the intersection of significance spaces depends in large part on the commensurability or incommensurability of the significance spaces involved. Specifically the base sixteen is in the binary family, and every hexadecimal number may be easily converted to binary by writing each hexadecimal digit as the appropriate four bits. For $b \epsilon S_{16}^6$, the 24 bit binary representation of the six hexadecimal digits of b may have up to three leading zeros, so some 22 bit binary numbers may not be representable in S_{16}^6, e.g. $(1+2^{-21}) \notin S_{16}^6$, however every 21 bit binary number will be contained in S_{16}^6. In generalizing this notion it is evident that if $\beta=\delta^p$, then each base β digit may be represented by p base δ digits. Thus an n-digit base β integer yields an np-digit base δ representation of which no more than p - 1 leading digits may be zero. Setting m = (n - 1)p + 1, we then have $S_\delta^m \subset S_\beta^n$, $S_\delta^{m+1} \not\subset S_\beta^n$, and the following theorem is derived.

Theorem 10: Let $\beta_1, \beta_2, \ldots, \beta_k$ be commensurable bases of the δ family. Then for $m = 1 + \min_i \{(n_i - 1)\log_\delta \beta_i\}$,

$$S_\delta^m \subset \bigcap_{i=1}^{k} S_{\beta_i}^{n_i} \text{ and } S_\delta^{m+1} \not\subset \bigcap_{i=1}^{k} S_{\beta_i}^{n_i} \tag{16}$$

Actually theorem 10 can be sharpened if we also consider the different intervals $[\beta^j, \beta^{j+1}]$. Reasoning as above it can be shown that by letting $p_i = \log_\delta \beta_i$, and setting $m = 1+\min_i \{p_i(n_i - 1) + (j \bmod p_i)\}$, then we have

$$S_\delta^m \cap [\beta^j, \beta^{j+1}] = \bigcap_{i=1}^{k} S_{\beta_i}^{n_i} \cap [\beta^j, \beta^{j+1}] \tag{17}$$

Furthermore if q is the minimizing value of i in the above definition of m, then also $S_\delta^m = S_{\beta_q}^{n_q}$ over the interval $[\beta^j, \beta^{j+1}]$.

These observations along with theorem 9 yield important properties of certain compound conversions. If Q is a compound truncation conversion through commensurable significance spaces, then $Q(x) \epsilon f(Q)$ for all real x. Hence QQ = Q, and if Q* is any compound truncation conversion through the same commensurable significance spaces as Q but in any order, then Q* = Q. Furthermore when $Q = T_{\beta_1}^{n_1} \ldots T_{\beta_k}^{n_k}$ with $\beta_1, \ldots, \beta_k$ in the same commensurable family, then $|x - Q(x)|/x < \max_i \Gamma_{\beta_i}^{n_i}(x)$, so accumulated conversion error is effectively controlled.

Now it will be shown that the intersection of incommensurable significance spaces does not contain any common significance space, for such an intersection has only a ***finite number*** of elements.

Theorem 11: If β and δ are incommensurable, then $S_\beta^n \cap S_\delta^m$ has no more than $2(\beta^n - 1)(\delta^m - 1)+1$ members.

Proof: Let P be the following set of ordered pairs of integers, $P = \{(k, k^*) \mid |k|<\beta^n, |k^*|<\delta^m \text{ and } kk^* \geq 1 \text{ or } k=k^*=0\}$. Define a mapping $P: S_\beta^n \cap S_\delta^m \to P$ as follows. P(0)=(0,0). Suppose $b \epsilon S_\beta^n \cap S_\delta^m$, $b \neq 0$. Then $b = k\beta^j$ where k and j may be chosen uniquely such that $|k| < \beta^n$ and β does not divide k. Similarly, there are unique k*, j* such that $b = k^*\delta^{j^*}$ where $|k^*| <\delta^m$ and δ does not divide k*. In this case P(b)=(k,k*). Thus P yields the nomalized (right shifted) integer portions of the representations of any element common to S_β^n and S_δ^m. It is now shown that P is a one-to-one mapping of $S_\beta^n \cap S_\delta^m$ into P.

Clearly only zero is mapped into (0,0), so let a, $b \epsilon S_\beta^n \cap S_\delta^m$ be any two non-zero elements where P(a) = P(b). Then there must exist k,k* $\neq 0$, and j, j*, i, i* such that $a = k\beta^j = k^*\delta^{j^*}$, and $b = k\beta^i = k^*\delta^{i^*}$ where β does not divide k and δ does not divide k*. Then $k/k^* = \delta^{j^*}/\beta^j = \delta^{i^*}/\beta^i$, so that $\delta^{j^*-i^*} = \beta^{j-i}$. Then j* = i* and j = i since β and δ are

incommensurable, so a=b and P is one-to-one. Now $S_\beta^n \cap S_\delta^m$ can have no more members than P which has $2(\beta^n - 1)(\delta^m - 1)+1$ elements, proving the theorem.

Utilizing the fundamental theorem of arithmetic (unique prime decomposition), much more can be said about the members of $S_\beta^n \cap S_\delta^m$ for particular β and δ. If β and δ are relatively prime, then all members of $S_\beta^n \cap S_\delta^m$ are integers, with the largest such integer strictly less than $\beta^n\delta^m$. If the greatest common divisor of the incommensurable β and δ is a prime number (as in the binary-decimal case), then the smallest positive element of $S_\beta^n \cap S_\delta^m$ can be shown to be a negative power of that prime. This smallest positive number exactly representable in both systems can be considerably larger than the underflow bound on a typical computer, for example $2^{-10} = .0009765625$ is the smallest positive member of $S_2^{24} \cap S_{10}^7$.

For a compound truncation conversion Q through $S_{\beta_i}^{n_i}$, i = 1, . . . , k, we have shown that Q(x) = x if and only if x is in the intersection of all significance spaces. Furthermore if at least one pair of the β_i are not commensurable, then the intersection is finite and $Q(x) \neq x$ for any positive x lower than the minimum positive element of the intersection. For such an x the compound truncation conversion Q would have $Q(x) > QQ(x) > QQQ(x) > \ldots > Q^{(j)}(x) > \ldots$, and in fact $\lim_{j\to\infty} Q^{(j)}(x) = 0$ since zero is the only finite accumulation point of $S_{\beta_k}^{n_k}$. Thus for all i, $Q^i \neq Q^{i-1}$ for any compound truncation conversion Q involving at least two incommensurable bases, in sharp contrast to a compound truncation conversion through commensurable significance spaces where the iterated conversions immediately converged. Practically speaking, the successive updatings of a B.C.D. tape[4] on a binary machine could cause some of the "constant" floating point data to be iteratively converted back-and-forth with each updating. If truncation conversion were adherred to as a standard, some of this data could drift lower in value (see figure 8), losing all accuracy, with no error indication provided by the system. Thus truncation conversion should be avoided in mixed base computation unless all bases are in the same commensurable family.

With rounding conversion the error accumulation upon successive conversions of a datum amongst incommensurable as well as commensurable bases is much better controlled. For example, it has been shown[8,3] that with suitably high significance in the intermediate space S_δ^m, the 2-fold compound conversions $R_\beta^n R_\delta^m$ and $R_\beta^n T_\delta^m$ can both reduce to the identity on S_β^n.

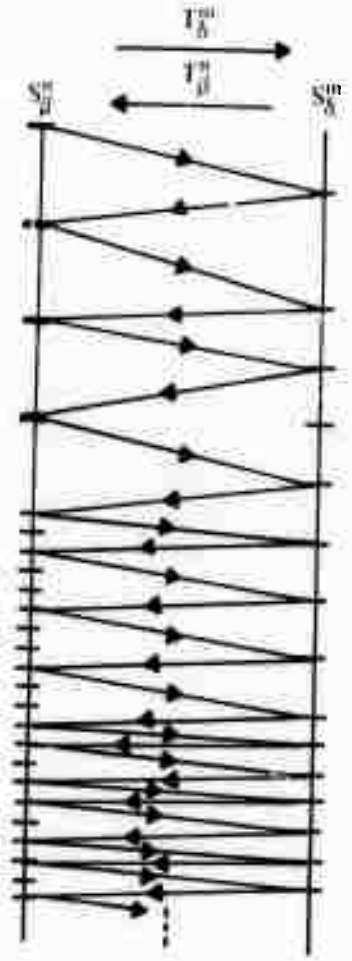

Figure 8: The possible drift in value of a "constant datum" under iterated truncation conversion between incommensurable significance spaces.

Theorem 12 (In-and-Out Conversion Theorem). For β and δ incommensurable,

i) $R_\beta^n R_\delta^m$ is the identity on $S_\beta^n \Leftrightarrow \delta^{m-1} > \beta^n$

ii) $R_\beta^n T_\delta^m$ is the identity on $S_\beta^n \Leftrightarrow \delta^{m-1} \geq 2\beta^n - 1$

Thus from theorem 12 we see that for certain compound conversions the invariant set will be the whole image space.

Corollary 12.1: For β and δ incommensurable,

$$\begin{aligned} &\text{i) } f(R_\beta^n R_\delta^m) = S_\beta^n \Leftrightarrow \delta^{m-1} > \beta^n \\ &\text{ii) } f(R_\beta^n T_\delta^m) = S_\beta^n \Leftrightarrow \delta^{m-1} \geq 2\beta^n - 1 \end{aligned} \tag{18}$$

The In-and-Out Conversion Theorem[3] has been stated for the case of incommensurable bases, however, incorporating the techniques of theorem 10, a more general result encompassing both commensurable and incommensurable bases can be conveniently derived for the 2-fold compound rounding conversion case.

Corollary 12.2: For $\beta, \delta \geq 2$, let γ be the greatest common root of β and δ when β and δ are commensurable, and let $\gamma = 1$ otherwise. Then

$$R_\beta^n R_\delta^m \text{ is the identity on } S_\beta^n \Leftrightarrow \gamma\delta^{m-1} \geq \beta^n \tag{19}$$

Proof: When β and δ are incommensurable $\gamma = 1$, and $\delta^{m-1} \neq \beta^n$, so (19) follows from theorem 12. Otherwise, when β and δ are commensurable with greatest common root γ, then $\delta = \gamma^i$, $\beta = \gamma^j$ where i and j are relatively prime. Hence $S_\beta^n \subset S_\gamma^{jn}$ and $S_\gamma^{(m-1)i+1} \subset S_\delta^m$. Now $\gamma\delta^{m-1} \geq \beta^n$ implies that $(m-1)i+1 \geq jn$, so $S_\beta^n \subset S_\delta^m$ and $R_\beta^n R_\delta^m$ is clearly the identity on S_β^n. Alternatively assuming $\gamma\delta^{m-1} < \beta^n$ means that $(m-1)i+1 < jn$. Now since i and j are relatively prime, there exists a k such that $k = j - 1 \bmod j$, $k = 0 \bmod i$. But then $S_\beta^n = S_\gamma^{jn}$ over the interval $[\gamma^k, \gamma^{k+1}]$ and also $S_\delta^m = S_\gamma^{(m-1)i+1}$ over the same interval, so that $(m-1)i+1 < jn$ means S_δ^m is contained in but not equal to S_β^n over the interval $[\gamma^k, \gamma^{k+1}]$. Thus $R_\beta^n R_\delta^m$ can not be the identity on S_β^n, completing the corollary.

When the condition $\gamma\delta^{m-1} \geq \beta^n$ is not obtained, then clearly $S_\beta^n \cap S_\delta^m \subset f(R_\beta^n R_\delta^m) \subsetneq S_\beta^n$, and it is of interest to give some alternative properties sufficient to characterize an invariant point of the compound conversion $R_\beta^n R_\delta^m$.

In general if the image space of the compound conversion Q were exactly equal to $f(Q)$, then $QQ=Q$, and a desirable situation controlling accumulated error is obtained. If the k-fold compound conversion Q ends with a truncation conversion, we have shown that QQ may not equal Q. Even if the final conversion of such a Q is a rounding conversion, the image space of Q can contain some points which are not invariant points of Q. For example, let $Q = R_3^2 R_2^4$. Since $2^{4-1} = 8 = 3^2 - 1$, by theorem 4 the mapping $R_3^2 | S_2^4 \rightarrow S_3^2$ is onto, so Q covers all of S_3^2. However from corollary 12.1, $f(R_3^2 R_2^4) \neq S_3^2$.

Some detailed criteria for determining invariant points of $R_\beta^n R_\delta^m$ will now be considered.

Lemma 13: Assume that $b \epsilon S_\beta^n$, $b \neq \pm\beta^i$ for any i, and for some $d \epsilon S_\delta^m$, $R_\beta^n(d) = b$. Then $R_\beta^n R_\delta^m(b) = b$, i.e. $b \epsilon f(R_\beta^n R_\delta^m)$.

Proof: Let β and δ be commensurable of the γ family. Let $b \epsilon S_\beta^n$, $b \neq \pm\beta^i$ for any i and assume $\gamma^j \leq |b| < \gamma^{j+1}$. Over the intervals $[\gamma^j, \gamma^{j+1}]$ and $[-\gamma^j, -\gamma^{j+1}]$ either $S_\beta^n \subset S_\delta^m$ or $S_\delta^m \subset S_\beta^n$, and in either case with $d \epsilon S_\delta^m$, $R_\beta^n(d) = b$ implies that $b \epsilon S_\delta^m \cap S_\beta^n$. Therefore $b \in f(R_\beta^n R_\delta^m)$.

Alternatively let us assume that β and δ are incommensurable. Let $0 < b' \epsilon S_\beta^n$ have predecessor b and successor b'', and assume that $b' \neq \beta^i$ for any i. Then $b'' - b' = b' - b$. We assume there exists a $d \epsilon S_\delta^m$ such that $R_\beta^n(d) = b'$, so

$$|d - b'| \leq (b'' - b)/2 \tag{20}$$

From the definition of rounding (10), it is evident that $|x - R_\delta^m(x)| = \min_{a \epsilon S_\delta^m} \{|x - a|\}$. Therefore with $x = b'$

$$|b' - R_\delta^m(b')| \leq |b' - d| \tag{21}$$

and again with $x = R_\delta^m(b')$

$$|R_\delta^m(b') - R_\beta^n R_\delta^m(b')| \leq |R_\delta^m(b') - b'| \tag{22}$$

and finally combining (20 - 22)

$$|b' - R_\beta^n R_\delta^m(b')| \leq b' - b \tag{23}$$

If (23) were an equality, then $d \neq R_\delta^m(b')$ and (20 - 22) would all be equalities. Therefore $|d - R_\delta^m(b')| = b' - b$, and furthermore there could be no members of S_δ^m between d and $R_\delta^m(b')$. Also then

$$b' = (d + R_\delta^m(b'))/2 \tag{24}$$

so that $R_\delta^m(b')$ would be the successor of d in S_δ^m. Now any element and its successor in S_δ^m must differ by an amount δ^j for some j. Similarly $b' - b = \beta^i$ for some i, so that with equality in (23) $\delta^j = \beta^i$. But then $i = j = 0$ since δ and β have been assumed incommensurable. Now $b' - b = \beta^0 = 1$ only if b and b' are consecutive integers, and similarly d and $d' = R_\delta^m(b')$ would be consecutive integers, contradicting (24). Therefore (23) must be a strict inequality, so that $R_\beta^n R_\delta^m(b')$ must equal b'. The case for negative $b' \epsilon S_\beta^n$ follows from sign complementarity, completing the lemma.

As a consequence of this lemma it is now shown that a comparison of the gap functions Γ_β^n and Γ_δ^m will suffice to determine many of the invariant points of $R_\beta^n R_\delta^m$.

Corollary 13.1: Assume $b \epsilon S_\beta^n$, $b \neq \pm\beta^i$ for any i, and $\Gamma_\beta^n(b) \geq \Gamma_\delta^m(b)$. Then $b \epsilon f(R_\beta^n R_\delta^m)$.

Proof: If $0 < b' \epsilon S_\beta^n$ with $b' \neq \beta^i$ for any i, then $b' - b = b'' - b'$, and the interval mapping into b' under R_β^n is the half open-half closed interval $[(b' - b)/2, (b'' - b')/2)$. Suppose no member of S_δ^m falls in this interval, then

$$\max\{d \mid d < b', d \epsilon S_\delta^m\} < (b' - b)/2$$

$$\min\{d \mid d \geq b', d \epsilon S_\delta^m\} \geq (b'' - b')/2$$

so that evaluating Γ_δ^m at the real number b' (see definition (6))

$$\Gamma_\delta^m(b') > (b'' - b')/b' = \Gamma_\beta^n(b')$$

Thus if $\Gamma^n_\beta(b') \geq \Gamma^m_\delta(b')$, then some member of S^m_δ must map into b' under R^n_β, and by lemma 13 $b' \epsilon f(R^n_\beta R^m_\delta)$. The result for negative $b \epsilon S^n_\beta$, $b \neq -\beta^i$, then follows from sign complementarity.

Thus it is readily possible to find sequences of invariant points of $R^n_\beta R^m_\delta$ from a gap function comparison.

Corollary 13.2: For $\beta^i < \delta^j$, assume that $\Gamma^n_\beta \geq \Gamma^m_\delta$ over the open interval $(\beta^i, \min\{\beta^{i+1}, \delta^j\})$. Then

$$(\beta^i, \min\{\beta^{i+1}, \delta^j\}) \cap S^n_\beta \subset f(R^n_\beta R^m_\delta)$$

Referring back to figure 2, it is evident from corollary 13.2 that all members of S^4_{16} greater than .0625 and less than 0.1 are invariant points of $R^4_{16}R^4_{10}$. Generally not all members of S^n_β from an interval, I, where $\Gamma^n_\beta < \Gamma^m_\delta$, will be members of $f(R^n_\beta R^m_\delta)$, but certainly any point of $R^n_\beta(S^m_\delta)$ other than a power of β must be an invariant point of $R^n_\beta R^m_\delta$. Since R^n_β restricted to $S^m_\delta \cap I$ is one-to-one to S^n_β, we may surmise that the number of members of $f(R^n_\beta R^m_\delta)$ in a neighborhood of x is comparable to the lesser of the numbers of members of S^n_β and of S^m_δ in that neighborhood. Hence although the members of $f(R^n_\beta R^m_\delta)$ are more erratically spaced, the relative difference between neighboring points of $f(R^n_\beta R^m_\delta)$ will still be bounded with a bound larger but not by an order of magnitude than the worst case in S^n_β and S^m_δ.

The preceding lemma and its corollaries do not quite provide the full story about $f(R^n_\beta R^m_\delta)$, since the integral powers of β and δ represent break points in their respective gap functions and by the preceding theory they must be treated separately. We have already pointed out that the image space of Q need not be identical to $f(Q)$ even for $Q = R^n_\beta R^m_\delta$, so the question of whether the iterates of Q, namely Q, QQ, QQQ, . . . , $Q^{(k)}$, . . . , will converge is still unanswered for $Q = R^n_\beta R^m_\delta$. This question is a very practical one since we would like to know if iterated rounding conversion between binary and decimal based systems will allow indefinite drift in the value of a "constant" as did truncation conversion, or if a stable pair of values must be achieved after a fixed number of rounding conversions back and forth. The following theorem is a surprisingly general and reassuring answer to this question.

Theorem 14: (Iterated Conversion Theorem): Let $Q = R^m_\delta R^n_\beta$ where $m, n \geq 2$. Then QQQ=QQ. Furthermore this result is best possible in the sense that there exists $\beta, \delta, n, m \geq 2$ such that $R^n_\beta QQ \neq R^n_\beta Q$.

Proof: For any real x, $R^m_\delta R^n_\beta R^m_\delta R^n_\beta(x) = R^m_\delta R^n_\beta(x)$ unless $R^m_\delta R^n_\beta(x) = \pm\delta^i$ for some i by lemma 13. Similarly with $Q = R^m_\delta R^n_\beta$, QQQ(x) = QQ(x) unless $QQ(x) = \pm\delta^j$ with $j \neq i$. Assuming $QQQ(x) \neq QQ(x)$, the definition of rounding conversion assures that

$$|R^m_\delta R^n_\beta(x) - R^n_\beta(x)| \geq |R^n_\beta R^m_\delta R^n_\beta(x) - R^m_\delta R^n_\beta(x)|$$

$$\geq |R^m_\delta R^n_\beta R^m_\delta R^n_\beta(x) - R^n_\beta R^m_\delta R^n_\beta(x)|$$

so that with $|Q(x)| = \delta^i$, $|QQ(x)| = \delta^j \neq \delta^i$,

$$|QQ(x) - Q(x)| \leqslant 2|R_\delta^m R_\beta^n(x) \quad R_\beta^n(x)| < \delta^i(1+\delta^{1-m}) - \delta^i$$

and also

$$|QQ(x) - Q(x)| = |\delta^i - \delta^j| = \delta^i|1 - \delta^{j-i}|$$

and since $m \geqslant 2$ by assumption in the theorem

$$|1 - \delta^{j-i}| < \delta^{1-m} \leqslant 1/\delta \leqslant 1/2$$

Now $|1 - \delta^{j-i}|$ has a positive integer value for $j > i$, and for $j < i$, the smallest value of $|1 - \delta^{j-i}|$ is $1 - 1/\delta \geqslant 1/2$, and this is a contradiction. Hence $QQ(x) = Q(x)$ for all x.

The remainder of the theorem demonstrating that $R_\beta^n QQ \neq R_\beta^n Q$ for some $n,m,\beta,\delta \geqslant 2$ is shown in the example of figure 9.

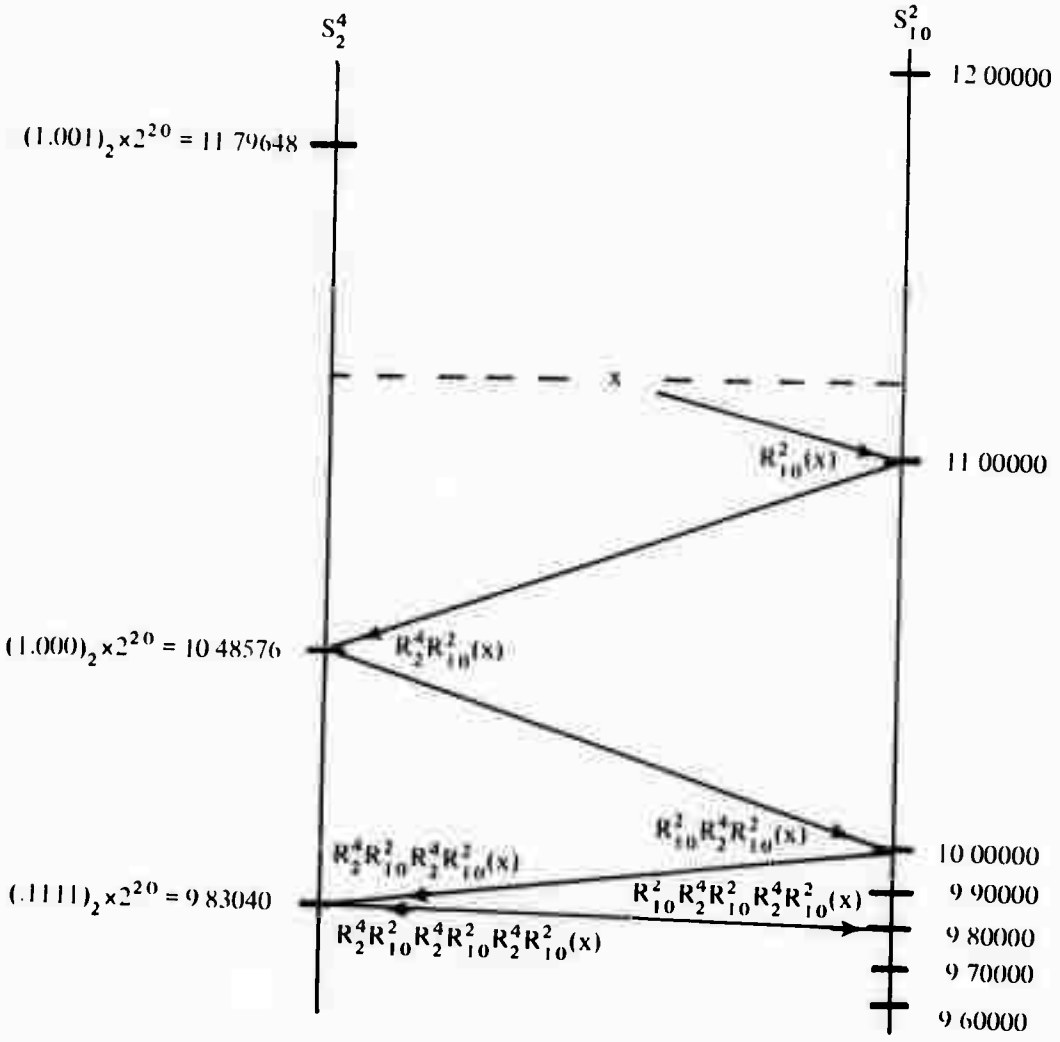

Figure 9: Iterated conversions of x = 1,120,000 by the compound conversion $R_2^4R_{10}^2$, showing that $R_{10}^2R_2^4R_{10}^2R_2^4R_{10}^2 \neq R_{10}^2R_2^4R_{10}^2$.

Computer hardware involving the bases 2, 8, and 16 of the binary family and the base 10 is in popular usage. Our results so far discuss accumulated error under compound conversion (1) within a commensurable family and (2) between two incommensurable significance spaces. The modification of our results to cover compound conversion between more than two significance spaces all from just two different commensurable families is straightforward, so that our theory does cover the current situations where one might expect to encounter mixed base computation.

If a suitable 3-state device were perfected for use in computer hardware, a ternary based floating point system would undoubtedly be implemented and the possibility of mixed base computation amongst the bases 2, 3 and 10 would ensue. Our final thoughts will then be concerned with compound conversion through a collection of incommensurable significance spaces

For the significance spaces S_{11}^5, S_2^{14}, and S_5^7, we have 1,960,563 = $(11\ 1A000)_{11} \in S_{11}^5$, 1,960,576 = $(1\ 11011\ 11010\ 10100\ 00000)_2 \in S_2^{14}$, and 1,960,500 = $(10002\ 14000)_5 \in S_5^7$, and the absolute difference between an element and its successor in this neighborhood is 121, 128, and 125 respectively (see figure 10). Thus the iterates of the compound rounding conversion $Q = R_5^7 R_2^{14} R_{11}^5$ in this neighborhood will be different for a number of cycles. As seen in figure 10, $Q^{(7)}(1{,}960{,}563) = 1{,}961{,}375$, and $Q^{(k)}(1{,}960{,}563) = 1{,}961{,}500$ for $k \geq 8$, hence $Q^{(k)} \neq Q^{(k-1)}$ for at least all $k \leq 8$. Examples such as this one show that utilizing rounding conversion is not enough to successfully restrict the drift in value of a constant datum under compound conversion. From the generalized result on In-and-Out Conversion given in corollary 12.2, a resolution to the problem of controlling overall error growth in the presence of more than two incommensurable bases by intermediate reconversions to a standard significance space is feasible.

Specifically let $S_{\beta_i}^{n_i}$, $1 \leq i \leq k$, be a collection of significance spaces representing the different floating point data formats of a mixed base computational environment. Suppose we introduce an intermediate space, S_δ^m, with the significance m small enough such that $R_\delta^m R_{\beta_i}^{n_i}$ is the identity on S_δ^m for all i. Then let all data introduced into the mixed base computational environment first be converted by rounding to S_δ^m, and let subsequent conversions from $S_{\beta_i}^{n_i}$ to $S_{\beta_j}^{n_j}$ be preceded by a reconversion to S_δ^m, (i.e. $R_{\beta_j}^{n_j} R_\delta^m | S_{\beta_i}^{n_i} \rightarrow S_{\beta_j}^{n_j}$). Note that the conversion from $S_{\beta_i}^{n_i}$ to S_δ^m always regenerates the same value, $R_\delta^m(x)$, in S_δ^m for an initial datum x, since $R_\delta^m R_{\beta_i}^{n_i}$ is the identity on S_δ^m. Thus the value of the initial datum x whenever encountered in $S_{\beta_i}^{n_i}$, even after numerous intermediate conversions is given by $R_{\beta_i}^{n_i} R_\delta^m(x)$, and this standardization of a constant's value with regards to each $S_{\beta_i}^{n_i}$ provides a highly desirable property for mixed base computation.

Now the range of possible values achievable in $S_{\beta_i}^{n_i}$ is $R_{\beta_i}^{n_i}(S_\delta^m)$, and it is of course desirable to have as large an m as possible so that the conversion through S_δ^m does not introduce too much error. Yet it should be kept in mind that if m is chosen so large that one of the $R_\delta^m R_{\beta_i}^{n_i}$ is not the identity on S_δ^m, then conversion error may accumulate and generate a greater overall error than if a smaller m (meaning less initial accuracy) were chosen. These observations can be formalized as an additional corollary to theorem 12 and corollary 12.2.

Corollary 12.3: For $S_{\beta_i}^{n_i}$, $1 \leq i \leq k$, let γ_i be the greatest common root of δ and β_i when they are commensurable, and let γ_i be unity otherwise. Let

$$m \leq \min_i \left\{ (n_i - 1) \log_\delta \beta_i + \log_\delta \gamma_i \right\} \tag{25}$$

and let $Q_i = R_{\beta_i}^{n_i} R_\delta^m$ for $1 \leq i \leq k$. If $Q = Q_j Q'$ for any $1 \leq j \leq k$ where Q' is composed from the mappings Q_i $1 \leq i \leq k$, then $Q = Q_j$.

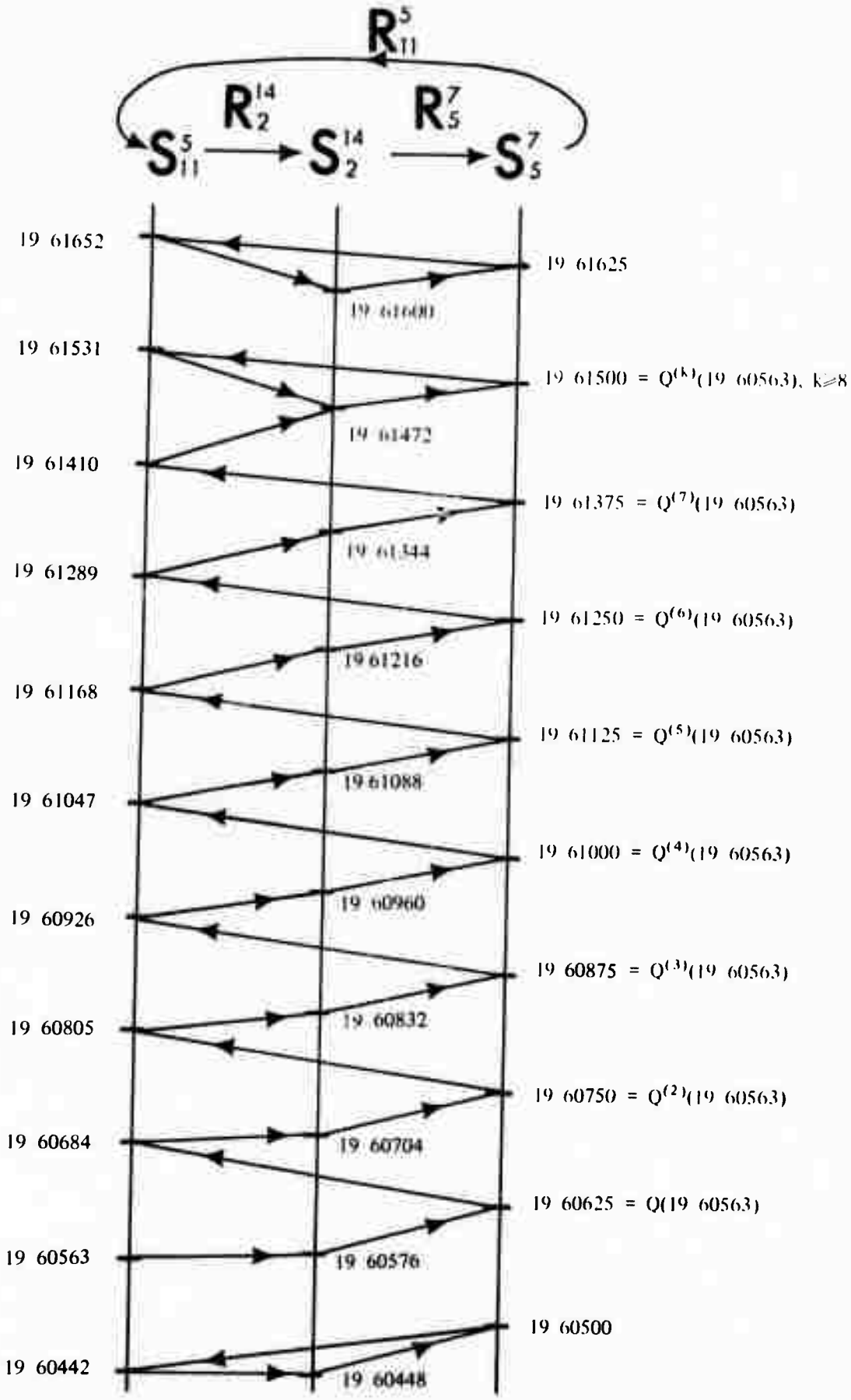

Figure 10. Iterates of the compound rounding conversion $Q = R_{11}^{5} R_{2}^{14} R_{5}^{7}$ showing the drift in value of a "constant datum" under successive conversions.

The fact that m must be bounded from above in (25) in order to guarantee control of accumulated error and avoid situations such as that exhibited in figure 10 demonstrates that the phrase "carry more digits" does not always mean that greater overall accuracy will follow, and such cliches should not be used as a substitute for a true understanding of the formal structure of floating point number systems and base conversion.

REFERENCES

1. Matula, D. W., "Base Conversion Mappings", *Proceedings of the Spring Joint Computer Conference, AFIPS*, Vol. 30, 1967, (311–318).

2. Matula, D. W., "The Base Conversion Theorem", *Proceedings of the American Mathematical Society*, Vol. 19, No. 3, June, 1968, (716–723).

3. Matula, D.W., "In-and-Out Conversions", *Communications of the Association for Computing Machinery*, Vol. 11, No. 1, 1968, (47–50).

4. Matula, D.W., "Towards an Abstract Mathematical Theory of Floating-Point Arithmetic", *Proceedings of the Spring Joint Computer Conference, AFIPS*, Vol. 34, 1969, (765–772).

5. Metropolis, N., Ashenhurst, R.L., "Radix Conversion in an Unnormalized Arithmetic System", *Mathematics of Computation*, Vol. 19, 1965, (435-441).

6. Collatz, Lothar, *Functional Analysis and Numerical Mathematics*, (translated by H. Oser), Academic Press, New York, 1966.

7. Urabe, M., "Roundoff Error Distribution in Fixed-Point Multiplication and a Remark About the Rounding Rule", *SIAM Journal of Numerical Analysis*, Vol. 5, No. 2, 1968, (202-210).

8. Hardy, G. H. and Wright, E.M., *An Introduction to the Theory of Numbers*, 3rd ed., Clarendon Press, Oxford, 1954. (See page 373)